



UK Research
and Innovation

HDRUK
Health Data Research UK



ADRUK
Data-driven change

DARE UK – Early Thinking

Core Federation Services

(formerly 'Digital Research Infrastructure')

Stakeholder workshop, Thursday 24 March 2022

Core Federation Services

The requirements discussed in this theme are central to delivering the core infrastructure elements of the DARE UK aims, which are to:

- Design and deliver a novel and innovative UK-wide data research infrastructure that is joined-up, demonstrates trustworthiness and supports research at scale for public good.
- Establish the next generation of trusted research environments (TREs) that will enable fast, safe and efficient sharing, linkage and advanced analysis of data, where it is legal and ethical to do so.
- Enable UK researchers and innovators to securely and efficiently harness the full power of linked datasets, modern digital platforms, tools, techniques and skills
- Enable research and analysis on a broad range of potentially sensitive data from across the UK research and innovation spectrum.

To enable efficient and trustworthy research on sensitive data, researchers require access to that data within a secure context that can support a wide range of analytical and computational capabilities. Trusted research environments (TREs) have emerged as the preferred solution for providing highly secure digital environments that enable remote access to information and analytical tools.

What is a TRE?

A TRE is a **Trusted Research Environment**. Also known as 'Data Safe Havens', TREs are highly secure computing environments that provide remote access to health data for approved researchers to use in research that can save and improve lives.

Why are they important?



TREs make research safer. Making data available through a TRE means that people can be **confident** that their personal health data is accessed **securely** and their **privacy protected**.

TREs help make research **efficient, collaborative** and **cost effective**, providing rich data that enables **deep insights** which will go on to improve healthcare and **save lives**.

TREs provide approved researchers with a **single location** to access valuable datasets. The data and analytical tools are all in **one place**, a bit like a **secure reference library**.

How is my data safeguarded?

Health data should always be kept safe and secure, and used responsibly to ensure privacy. Health Data Research UK ensures these high standards are met by promoting the use of the 'Five Safes' model across all TREs.

- 

Safe People
Only trained and specifically accredited researchers can access the data
- 

Safe Projects
Data is only used for ethical, approved research with the potential for clear public benefit
- 

Safe Settings
Access to data is only possible using secure technology systems – the data never leaves the TRE
- 

Safe Data
Researchers only use data that have been de-identified to protect privacy
- 

Safe Outputs
All research outputs are checked to ensure they cannot be used to identify subjects

First Draft Recommendations (1 of 4)

CFS Recommendation 1 Develop Reference Architecture for TREs

- 1a. Develop a reference architecture and open implementation for a “TRE in a Box” using open source technologies suitable for deployment on-premise or on public cloud, and that would accelerate the ability of existing infrastructure to move to a hybrid cloud model.
- 1b. Develop a reference architecture and open implementation for a “Pop-Up TRE” that can be deployed within existing TRE environments and alongside the “TRE in a Box” reference architecture to support secure transient analysis of data from multiple TREs.
- 1c. Investigate approaches to integrate the TRE network with the future Large-Scale Compute provisioning.
- 1d. Pilot a project to cloud-burst to high scale spot market instances on public cloud, with a view to potentially building a shared capability to provide this as a service across cloud providers in Phase 3.

Develop reference architecture for TREs

- Need to build this alongside standardised processes and governance, including accreditation processes. Pop-up TREs also need to meet data sharing agreements and governance processes.
- Not a silver bullet – teams will still need to know how to run a secure production environment.
- Would make integration easier as TREs would have standard external interfaces.
- Supportive of a formally managed open source approach, e.g. [Apache Software Foundation](#) and others.
- Could this be provided centrally as a ‘service’?
- If built on public cloud (i.e. via services provided over the internet by a third party), how do we ensure security – do we need to insist on UK provisioning?
- Needs to build from existing work – e.g. Turing, Open Source Azure project.
- Needs to ensure existing infrastructures are not excluded, and their progress not wasted; and that there is flexibility, as one size doesn’t fit all.
- Needs to be future proofed; there must be investment to evolve.

First Draft Recommendations (2 of 4)

CFS Recommendation 2 Establish approach to High Availability and Disaster Recovery

- 2a. Undertake a study to establish the business continuity and disaster recovery requirements for a production network of TREs.
- 2b. Pilot a network failover capability to support DR requirements.
- 2c. Develop a HA extension for the “TRE in a Box”.

Establish approach to High Availability and Disaster Recovery

- Mixed views on the value of investment in this area. Is the costs appropriate to risk? Needs further investigation before major investment.
- Could availability through public cloud (i.e. via services provided over the internet by a third party) offset risks?
- Limited attention paid to this beyond backup strategy currently.
- Needs to work with data protection legislation; important to assess the risk of moving data.
- Could this be centralised, e.g. via a data stewardship organisation which runs the service but has no access to unencrypted data?
- Costs of replication of data prohibitive for genomics, imaging, etc.
- Is there a need to identify “crown jewels”; is there some data that needs greater protection?

First Draft Recommendations (3 of 4)

CFS Recommendation 3 Build APIs to support Core Federation Services

- 3a. Design and assemble an open reference API library to support core federations services building on existing open source projects.
- 3b. Deploy the Federation API Library as a PoC with a driver usecase across three TREs from different research domains.
- 3c. Develop a Proof of Concept for a cloud-native implementation of a portable analytics workspace.
- 3d. Conduct a study to identify the requirements for a cross-domain data management and linkage service.

Build APIs (Application Programming Interfaces) to support Core Federation Services

- How do we ensure federation does not breach data sharing agreements?
- There needs to be more of a risk-based service; how do we segment data based on risk?
- Work with existing TREs – perhaps a pair – to implement services together.
- Don't make this centralised – needs to be community-led and owned. Alternative view was to have governance under a central organisation.
- Must be properly governed open source; no lock-in.
- There needs to be more focus on supporting federated analytics and machine learning APIs.
- Will need a reference implementation, and Proof of Concepts/examples to show how to use the APIs.
- Need to resolve where responsibility will rest if APIs don't work. How do we provide support?
- Need to ensure they support real science and not just be interesting papers about APIs.
- Look to assemble rather than reinvent – there are plenty of example projects already available; good use cases will be essential to the development of the right APIs.

First Draft Recommendations (4 of 4)

CFS Recommendation 4. Run a competitive call for driver projects to utilise the new infrastructure services and to validate that they are fit for purpose.

- 4a. Pilot a cross council usecase to validate the capabilities delivered in CFS Recommendation 3.
- 4b. Identify usecases to act as driver projects to validate the progressive rollout of production deployment in Phase 3.



Run a competitive call for driver projects

- Essential, but needs to be low risks to ensure participation.
- Shorter projects may be a challenge if the teams need to get access to sensitive data.
- Can this be tied into some of the existing innovation initiatives?
- Cover both essential and edge use cases.
- Brilliant real world use cases will demonstrate value far better than documents.



UK Research
and Innovation

HDRUK
Health Data Research UK



ADRUK
Data-driven change

Thank you

Find out more about DARE UK: www.dareuk.org.uk